Continual Self-Supervised Learning for Scalable Multi-script Handwritten Text Recognition

Marwa Dhiaf[1,3,4], Mohamed Ali Souibgui[2], Ahmed Cheikh Rouhou[1], Kai Wang[2], Yuyang Liu[2], Yousri Kessentini[3,4], Alicia Fornés[2],

[1]InstaDeep, [2]Computer Vision Center, Universitat Autònoma de Barcelona, [3]Digital Research Center of Sfax, [4]SM@RTS Tunisia

Abstract

We explore the potential of continual self-supervised learning to alleviate the catastrophic forgetting problem in handwritten text recognition, as an example of sequence recognition. Our method consists of adding intermediate layers called adapters for each task, and efficiently distilling knowledge from the previous model while learning the current task. Our proposed framework is efficient in both computation and memory complexity. To demonstrate its effectiveness, we evaluate our method by transferring the learned model to diverse text recognition downstream tasks, including Latin and non-Latin scripts.

Motivation

- Self Supervised Learning-based (SSL) models should be able to progressively and continuously learn new tasks without forgetting the previous ones, and also, whenever new unlabeled data is available.
- Retraining the model with the full dataset(old+new) is impractical, costly, and even impossible when previous data is not available anymore.
- Our work is the first application of continual learning to the field of handwritten text recognition, showing its ability to continuously learn from new scripts or languages.
- An efficient continual self-supervised learning framework in terms of complexity and storing memory is proposed to address the issue of catastrophic forgetting with no need for prior knowledge at inference time.
- Our approach addresses privacy concerns prevalent in various scenarios, where the storage of entire images containing sensitive information may not be feasible.

Datasets and metrics

Datasets:

Three examples of the images from the different languages/scripts datasets that were used for the experiments: IAM[3], LAM[4], and HKR[5].

Metrics:

For the evaluation, we use the Character Error Rate (CER) between the produced text output and the ground: truth.

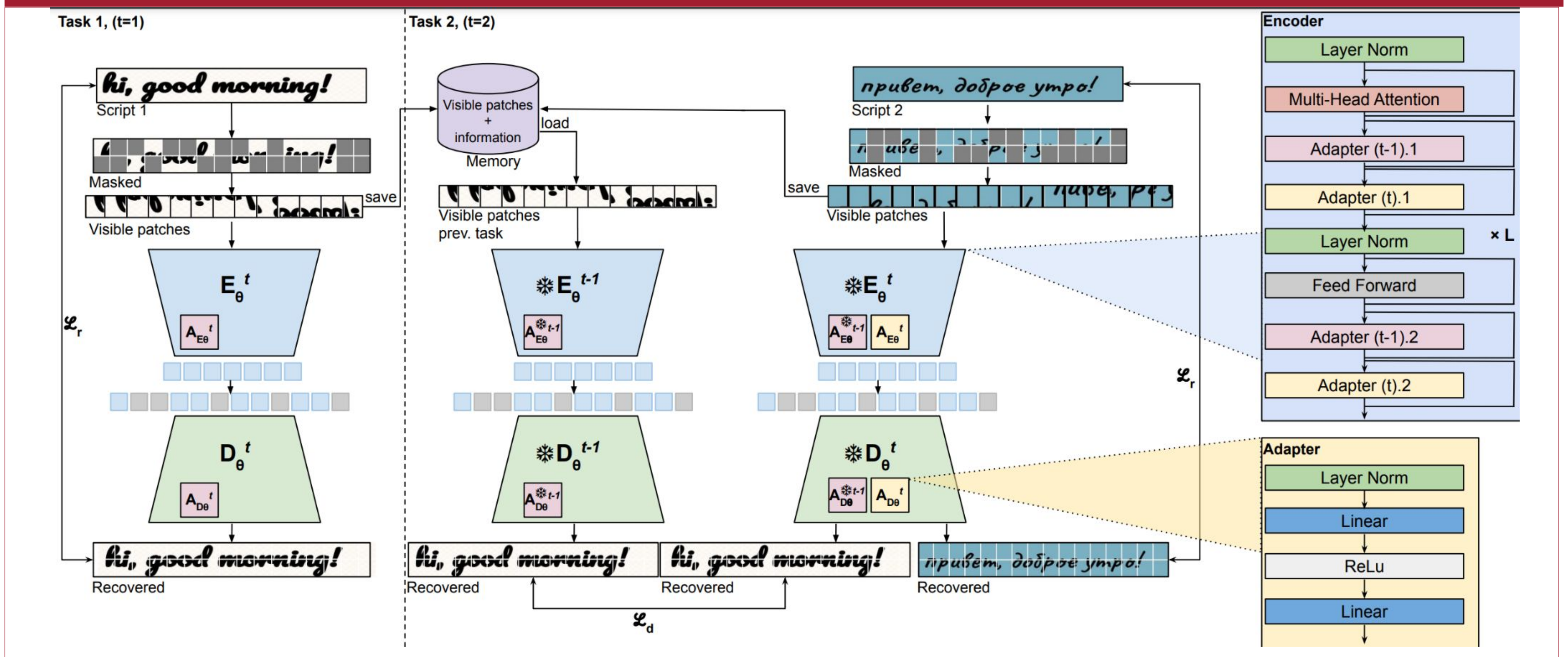
Implementation details:

ConFigure	Pre-Training	Fine-tuning
Optimizer	AdamW	Adam
Learning rate	1.5 e^{-4}	5 e^{-5}
Weight decay	0.05	0.05
Optimizer momentum	β_1, β_2=0.9, 0.95	β_1, β_2=0.9, 0.95
Batch size	32	32
Learning rate schedule	cosine decay	cosine decay
Warmup epochs	3	3

The recovery of the English dataset

GT (English):
Masked (t=1):
Recov. (t=1):
Masked (t=2):
Recov. (t=2):
Masked (t=3):
Recov. (t=3):
GT (Italian):
Masked (t=2):
Recov. (t=2):
Masked (t=3):
Recov. (t=3):

Architecture



Pre-training phase

At each time step t, only the task-specific adapter is learned and the rest of the model is frozen to overcome the catastrophic forgetting.

Also, a set of visual patches from the data is stored in a memory buffer, and others from previous tasks, are loaded and replayed into the current model for a task-agnostic inference. The proposed model is made up of an encoder-decoder module with task-specific adapter components.

- **Encoder** : our encoder is vanilla ViT [1] backbone.
- **Decoder**: The decoder is a transformer reconstruction module. As in MAE [2], the encoded tokens are first concatenated with a set of mask tokens that indicated the presence of the missing patches that should be predicted.
- **Adapter**: adapter components are designed to efficiently adapt the model for a new script/language. The idea is to embed them after each Mult-Head Attention (MHA) and feedforward layer within a transformer block.

Quantitative results

1. Evaluation of continually pre-trained models :

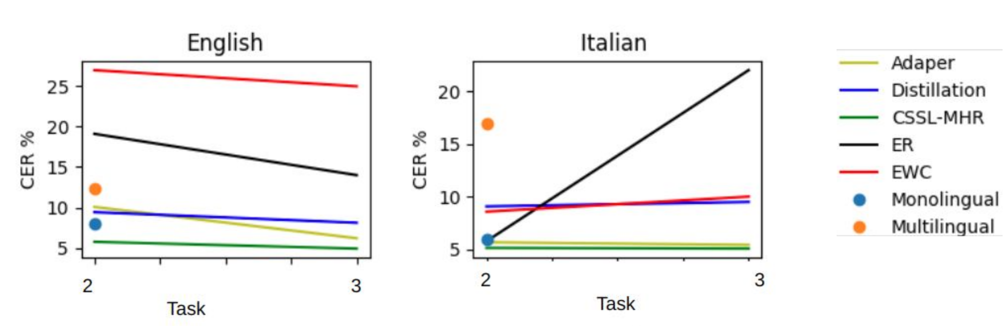
Method	FT after Task 2		FT after Task 3			# Tr. parms.	# Samples ↑
	English ↓	Italian ↓	English ↓	Italian ↓	Russian ↓		
Multilingual	12.3	17.0	12.3	17.0	4.7	68 M	–
Monolingual	8.0	6.0	8.0	6.0	1.4	68 M	–
Adapter	10.6	5.7	6.1	5.4	**2.8**	10.2 M	–
Distillation	9.4	9.1	8.1	9.5	6.6	68 M	3200
ER (Rolnick et al. 2019)	19.1	5.8	14.0	22.0	5.9	68 M	1920
EWC (Kirkpatrick et al. 2017)	27.0	8.6	25.9	10.0	5.7	68 M	–
CSSL-MHTR	**5.7**	**5.1**	**4.9**	**5.1**	2.9	**10.2** M	**3200**

2. Comparison with state-of-the-art approaches for HTR :

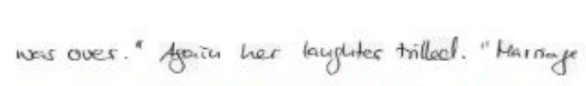
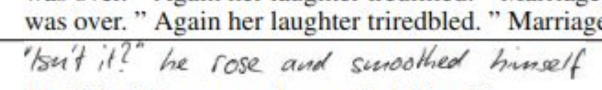
System	English	Italian	Russian
1D-LSTM(Puigcerver 2017)	8.3	3.7	69.1
CRNN (Cojocaru et al. 2021)	6.8	3.3	-
Transformer (Kang et al. 2022)	4.6	10.2	-
TrOCR (Li et al. 2021)	3.4	3.6	-
GFCN (Coquenet, Chatelain, and Paquet 2020)	8.0	5.2	-
OrigamiNet (Yousef and Bishop 2020)	4.8	**3.1**	-
Bluche (Bluche and Messina 2017)	**3.2**		22.3
G-CNN-BGRU (Abdallah, Hamada, and Nurseitov 2020)	-	-	8.3
CSSL-MHTR	4.9	5.1	**2.9**

Qualitative results

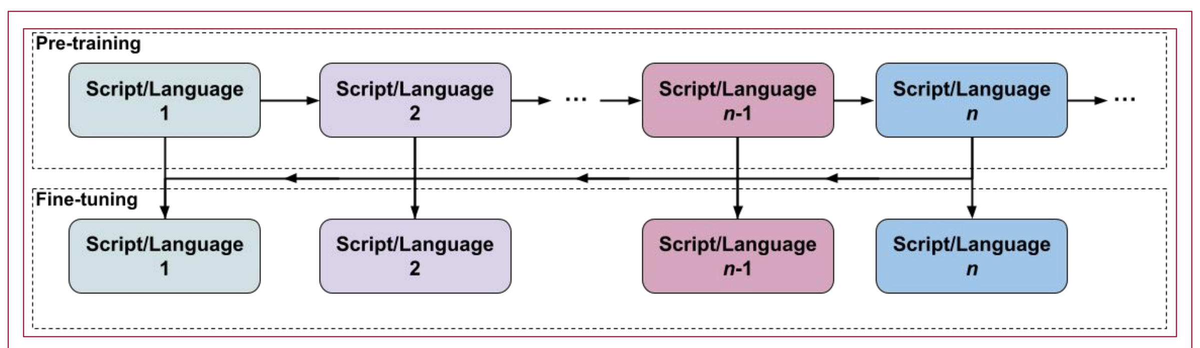
1. **The obtained CER after the continual pre-training :**



2. **The output of the different models when recognizing English text image :**

Input:	
GT:	was over. " Again her laughter trilled. " Marriage
Monolingual:	was over " Aredfein redler laredyghteredd trediiredbled. " Mredoxoyage
CSSL-MHTR (t=2):	was over. " Again her laughter trediilled. " Marriage
CSSL-MHTR (t=3):	was over. " Again her laughter triredbled. " Marriage
Input:	
GT:	" Isn't it ? " he rose and smoothed himself
Monolingual:	" redtsreda't it " " he rose and sreduoothed himself
CSSL-MHTR (t=2):	" Isn't it red' " he rose and smoothed himself
CSSL-MHTR (t=3):	" Isn't it ? " he rose and smoothed himself

Fine-tuning phase



Conclusion and Futures works

Our proposed CSSL-MHTR approach consists of an encoder-decoder transformer model that includes language/script adapter components and a memory replay strategy for continual self-supervised learning on handwritten text images.

As future work, we plan to extend this approach to recognize full pages instead of segmented lines. Also, we will explore the addition of other document analysis tasks to be learned continually, for instance, layout analysis, name entity recognition and information extraction.

References

[1] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai,X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.;Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

[2] He, K.; Chen, X.; Xie, S.; Li, Y.; Doll´ar, P.; and Girshick, R. 2022.Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000–16009.

[3] Marti, U.-V.; and Bunke, H. 2002. The IAM-database: an English sentence database for offline handwriting recognition.International Journal on Document Analysis and Recognition 5: 39–46.

[4] Cascianelli, S.; Pippi, V.; Maarand, M.; Cornia, M.; Baraldi, L.;Kermorvant, C.; and Cucchiara, R. 2022. The LAM Dataset: A Novel Benchmark for Line-Level Handwritten Text Recognition. In 2022 26th International Conference on Pattern Recognition(ICPR), 1506–1513. IEEE. Nurseitov, D.; Bostanbekov, K.; Kurmankhojayev, D.; Alimova.

[5] A.; Abdallah, A.; and Tolegenov, R. 2021.n Handwritten Kazakh and Russian (HKR) database for text recognition.
Multimedia Tools and Applications , 80: 33075–33097.

[6] Kessler, S.; Thomas, B.; and Karout, S. 2022. An adapter based pre-training for efficient and scalable self-supervised speech representation learning.
In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
,3179–3183. IEEE.

[7] Souibgui, M. A.; Biswas, S.; Mafla, A.; Biten, A. F.; Fornes, A.;Kessentini, Y.; Llad ́os, J.; Gomez, L.; and Karatzas, D. 2022a.Text-DIAE: A Self-Supervised Degradation Invariant Autoencoders for Text Recognition and Document Enhancement.arXiv preprint arXiv:2203.04814